Developing a Comprehensive U.S. Transit Accessibility Database

Andrew Owen (corresponding author)
University of Minnesota, Department of Civil, Environmental, and Geo-Engineering
500 Pillsbury Drive SE
Minneapolis, MN 55408
aowen@umn.edu

David M. Levinson
University of Minnesota, Department of Civil, Environmental, and Geo-Engineering
500 Pillsbury Drive SE
Minneapolis, MN 55408
dlevinson@umn.edu

Developing a Comprehensive US Transit Accessibility Database

Abstract

This paper discusses the development of a national public transit job accessibility evaluation framework, focusing on lessons learned, data source evaluation and selection, calculation methodology, and examples of accessibility evaluation results. The accessibility evaluation framework described here builds on methods developed in earlier projects, extended for use on a national scale and at the Census block level. Application on a national scale involves assembling and processing a comprehensive national database of public transit network topology and travel times. This database incorporates the computational advancement of calculating accessibility continuously for every minute within a departure time window of interest. This increases computational complexity, but provides a very robust representation of the interaction between transit service frequency and accessibility at multiple departure times.

Key Words

Accessibility, connectivity, transit

Introduction

Accessibility measures the number of opportunities that can be reached in a given travel time — an important metric for assessing the effectiveness of transportation–land use systems. To date, while these metrics have been used locally, there has been no standardized way to compare metropolitan areas systematically. This paper described the development of an integrated software framework for a nationwide evaluation of the accessibility to jobs provided by public transit systems at the Census block level. Application on a national scale involves assembling and processing a comprehensive national database of public transit network topology and travel times. This database incorporates the computational advancement of calculating accessibility continuously for every minute within a departure time window of interest. Values for contiguous departure time spans can then be averaged or analyzed for variance over time. This increases computational complexity, but provides a very robust representation of the interaction between transit service frequency and accessibility at multiple departure times.

This project focused on measuring access to jobs, and the output dataset indicates how many jobs can be reached from each Census block within various travel time thresholds, assuming trips made by walking and transit. With minor modifications, this framework can be adapted to provide accessibility metrics for any destination type.

The development of a comprehensive and consistent national public transit accessibility database involved three major components. First, appropriate data sources were identified, collected, and aggregated in a single input geodatabase. Second, a travel time calculation methodology was selected which provides a reasonable and useful representation of expected travel times by public transit. Finally, block-level travel times and the resulting accessibility were calculated in a parallelized, scalable cloud computing environment.

The following sections overview the project's motivation, goals, and implementation and discuss lessons learned and future directions for improving the research and practice of accessibility evaluation.

Motivation and Goals

In both practice and in research, accessibility evaluation remains experimental and methodologically fragmented: researchers and planners focusing on different geographical areas often implement different techniques, making it difficult to compare accessibility metrics across different locations. This encourages the

development and refinement, of improved accessibility evaluation techniques, but heightens the "first mover" risk for agencies seeking to implement accessibility-based planning practices, as they must select a method that might produce results that can only be interpreted locally. Development of a common baseline accessibility metrics advances the use of accessibility-based planning in two ways. First, it provides a stable target for agencies seeking to implement accessibility-based methods in upcoming planning processes. Second, it provides researchers a frame of reference against which new developments in accessibility evaluation can be compared.

In 2012, the Minnesota Department of Transportation (MnDOT) implemented an "Annual Accessibility Measure for the Twin Cities Metropolitan Area" that provides a methodology for calculating accessibility in the Minneapolis–Saint Paul metropolitan area, and that establishes an evaluation methodology for accessibility to jobs by car and transit (Owen & Levinson 2012). Development phases of this project relied on proprietary and custom transit schedule data formats because the GTFS format (described below) had not been adopted by local transit operates (Krizek et al. 2007, 2009).

Simultaneously, the value of consistent, systematic accessibility evaluations across multiple metropolitan areas was demonstrated by the work of Levine et al. (2012), which collected zone-to-zone travel time information from 38 metropolitan planning organizations to implement a cross-metropolitan evaluation of accessibility by car.

The goal of this project is to combine the lessons learned from these earlier works with recent advances in transit schedule data format and availability to produce a new, comprehensive dataset of accessibility to jobs by transit.

Data Sources

Detailed digital transit schedules in a consistent format are a critical component of this system, and the availability of such data is a relatively recent phenomenon. The General Transit Feed Specification (GTFS) (Google 2013) was developed by Google and Portland TriMet as a way to provide transit schedules for use in traveler routing and information tools.

Though the initial goal of GTFS was to provide a common format for traveler-focused schedule and routing software, it has also become a key resource for research and analysis of transit systems. Jariyasunant et al. (2011) and Delling et al. (2013) describe recent work in algorithmic approaches to calculating travel times on transit networks that rely on GTFS. Puchalsky et al. (2012) describe how the stop and schedule data contained in GTFS datasets can strengthen regional planning and forecasting processes. Wong (2013) examines how data currently available in GTFS

enables network- and agency-level analysis of transit systems, while Catala et al. (2011) identifies ways that the GTFS format could be expanded to support additional uses in transit operations and planning. It would be difficult to overstate the importance of the GTFS data format, and its widespread adoption, in enabling consistent analysis methodology across multiple transit operators.

Despite their importance and digital nature, the collection of GTFS datasets can be frustratingly inconsistent and error-prone. While the format of GTFS data itself is standardized there are no standards for the digital publication of the datasets, and practices vary widely across transit operators. A majority of operators (at least among medium and large metropolitan areas) provide GTFS datasets via a direct web site link. However, even among these variations in URL naming conventions pose challenges for systematic retrieval. Other operators allow GTFS dataset downloads only after users interactively submit a form or agreement. Still others generate GTFS datasets and provided them directly to Google. for use in their popular online routing tool, but release them to the public only in response to direct email or hard-copy requests.

These issues are somewhat mitigated by the web site www.gtfs-data-exchange.com, a crowd-sourced archive of GTFS datasets from around the world. However, the crowd-sourced nature of this resource poses its own challenges. Most importantly, it is difficult — and in some cases impossible — to validate that a GTFS dataset obtained from this source was originally published by the actual transit operate, or that it has not been modified in some way. For this project, schedules downloaded from this web site are used only when they cannot be obtained directly from a transit operator.

Software

All of the major components of this evaluation system are open source. While this was not a specific goal or requirement, experience from earlier projects suggested some important benefits of using open source tools. First, open source software often provided greater flexibility in input and output data formats. This is an important consideration when a project involves multiple stages of data transformation and processing, each performed with a separate tool. Second, open source software can be rapidly customized to fit the project needs. In this project, local customizations to OpenTripPlanner provided more efficient parallelization and allowed for better data interoperability. Finally, open source approaches reduce barriers to replication and validation. Because the output of this project is itself a dataset designed for use in research and practice, it is important that all parts of the methodology — including those implemented using existing software — are thoroughly transparent and understandable.

This project makes use of the following major software packages:

- **OpenTripPlanner** (OTP), an open-source platform for multi-modal journey planning and travel time calculation.
- **PostgreSQL**, an open-source SQL database engine.
- **PostGIS**, a PostgreSQL extension that allows efficient storage and querying of spatial data.

Additionally, numerous smaller scripts and tools for data collection and processing were developed specifically for this project.

Data Processing and Organization

Figure 1 illustrates the basic project architecture and workflow, which is described in the following sections.

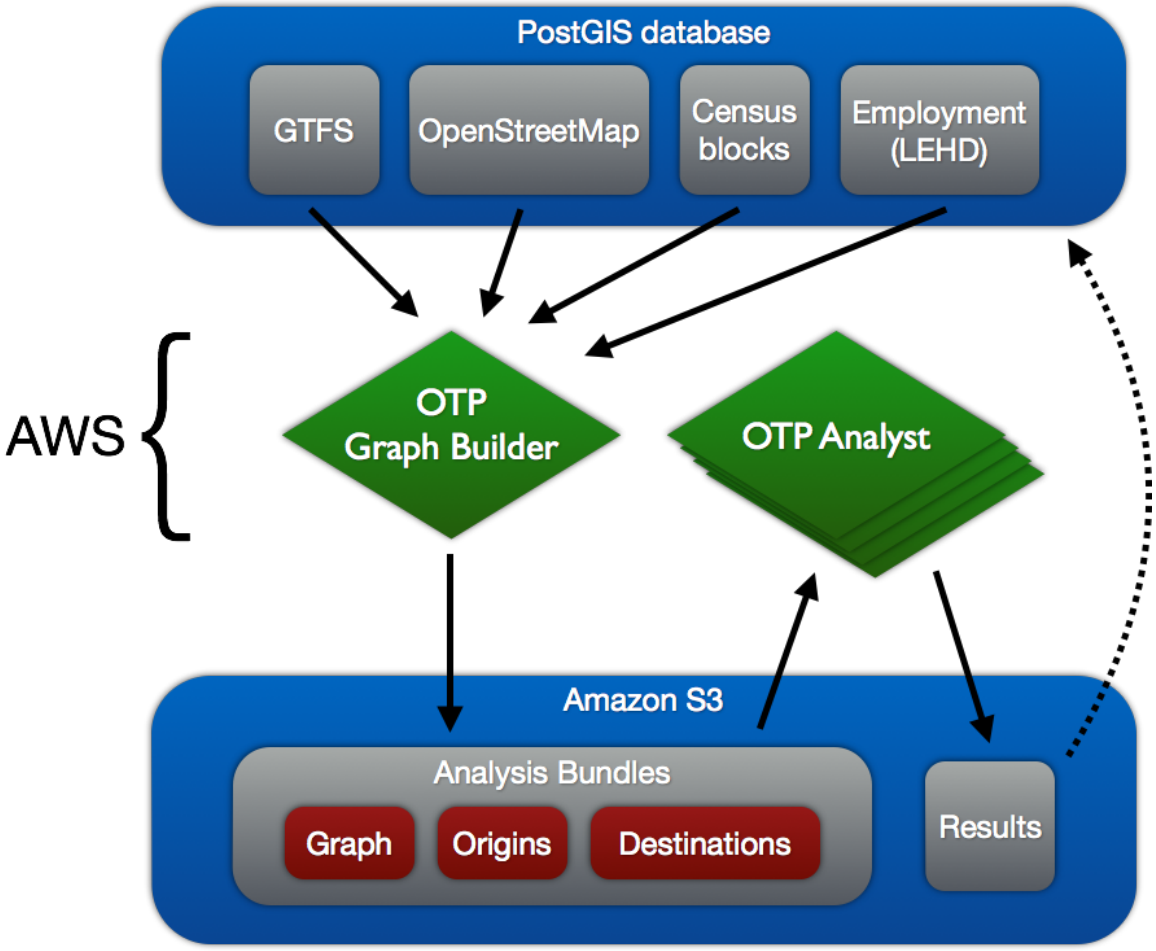

Figure 1: Project architecture and workflow

Inputs

The project inputs are stored primarily in a single SQL database. PostgreSQL is used along with the PostGIS extension; this combination allows spatial and non-spatial data in a single database, automated spatial queries (*e.g.* to select all origins within a

given analysis zone), and spatial indexing methods that accelerate these queries. Specifically this database contains an extract of all OpenStreetMap pedestrian data for North America; the full block, county, and core-based statistical area (CBSA) datasets from the U.S. Census Bureau; all 2011 resident area characteristics (RAC) and workplace area characteristics (WAC) data files from LEHD, and spatial bounds information for all collected GTFS datasets (which are stored separately).

Calculation

Travel time calculation is an "embarrassingly parallel" problem — a popular term among computer scientists for computation scenarios that can be easily decomposed into many independent repetitions of the same basic task. Given a suitable data architecture, these tasks can then be performed simultaneously, exponentially increasing the overall calculation speed.

In this case, the calculation of travel times form one origin at one departure time follows exactly the same process as for every other origin and every other departure time. Just under 11.1 million Census blocks (2010) comprise the United States; combined with 1,440 minutes in a day this gives almost 16 billion possible space-time origins. The effective number is less, however, because in block with no access to transit service only a single departure time is used — transit travel times vary significantly over the day but walking travel times do not.

The core unit of work — calculating travel times from a single origin at a single departure time — is provided by existing OpenTripPlanner capabilities. The parameters and assumptions involved in these calculations are described in following sections. OTP is natively multithreaded and can efficiently parallelize its work across multiple processors. To achieve efficient parallelization without requiring dedicated supercomputing techniques, the total computation workload is divided into "analysis bundles" which include all information necessary to compute a defined chunk of the final data. Each analysis bundle includes origin locations and IDs; destination locations, IDs, and opportunity (job) counts; and a unified pedestrian-transit network created by OTP.

The scope or origins included in each bundle is arbitrary; a useful value of 5,000 origins per bundle was found through trial and error. Figure 2 illustrates the division of a single county into analysis zones, each containing no more that 5,000 census block centroids. Too-small bundles erode overall efficiency by increasing the overhead costs of job tracking and data transfer, while too-big bundles suffer reliability issues: errors do occur, and when they do it is preferable to lose a small amount of completed work rather than a large amount.

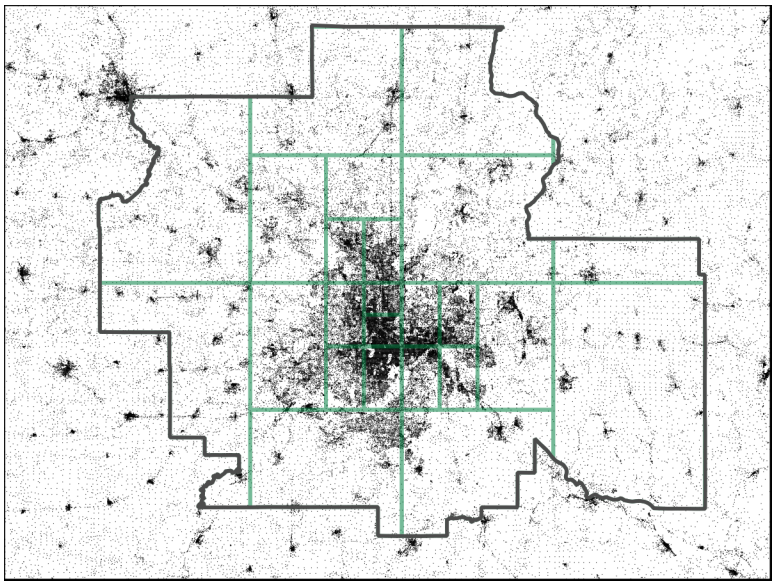

Figure 2: A metropolitan area divided into analysis zones. Each zone contains a maximum of 5,000 Census block centroids.

Destinations, on the other hand, are selected geographically. Because travel times are by definition not known until the calculations are complete, it is necessary to include in each bundle all destinations that might be reached from any of the included origins within some maximum time threshold. A buffer of 60 km from the border of the origin zone is used, based on 1 hour of travel at an estimated 60 km/h upper limit of the average speed of transit trips. This 1-hour limit only applies to the extent of the graph; using such a graph, accessibility metrics can be reported for any time threshold of 1 hour or less. Figure 3 illustrates the spatial selection of destinations for a given set of origins.

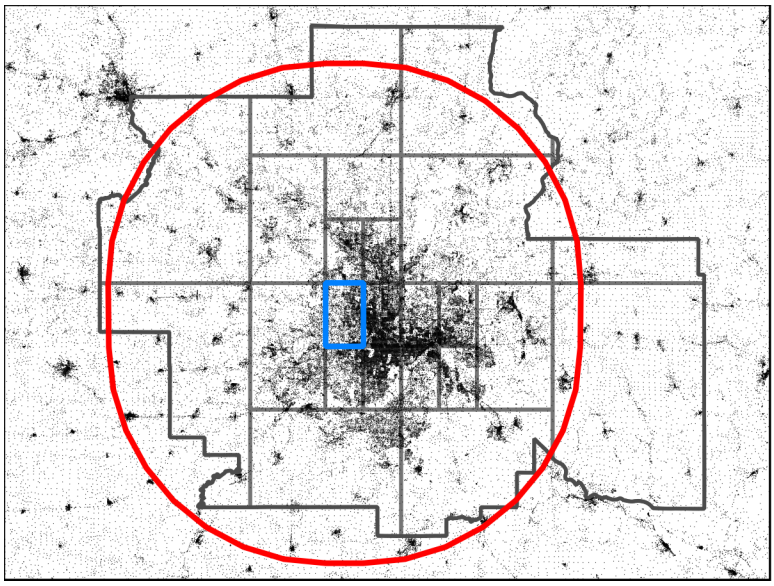

Figure 3: A single origin zone (blue) and its corresponding 60-kilometer destination zone buffer (red). Travel times are calculated from each centroid in the origin zone to each centroid in the destination zone.

OTP's Analyst module provides a graph builder function that combines pedestrian and transit network data from the input database into a single graph, and locally-developed software merges the graph into an analysis bundle with the appropriate origins and destinations. The bundle is queued in a cloud storage system making it available for computation.

Computations take place on a variable number of cloud computing nodes that are temporarily leased while calculation is in progress. (Currently, computing nodes are leased from Amazon Web Services (AWS).) Each node is prepared with OTP Analyst software as well as custom software that retrieves available analysis bundles, initiates accessibility calculations, and stores the results.

Outputs

The processing of each analysis bundle results in a single data file that records accessibility values for each origin in the bundle. For each origin, this includes an accessibility value for each departure time and for each travel threshold between 5 and 60 minutes, in 5-minute increments. These values are stored individually and disaggregated to facilitate a wide range of possible analyses. Each result file is tagged with the ID of the analysis zone and range of departure times for which it contains results, and then stored in a compressed format in the cloud storage system.

Because analysis typically takes place at the metropolitan level or smaller, it is rarely necessary to have the entire national result dataset available at once. Instead, custom scripts automate the download of relevant data from the cloud storage system.

Accessibility Calculations

Transit Travel Time

This analysis makes the assumption that all access portions of the trip — initial, transfer(s), and destination — take place by walking at a speed of 1.38 meters/second along designated pedestrian facilities such as sidewalks, trails, etc. On-vehicle travel time is derived directly from published transit timetables, under an assumption of perfect schedule adherence. Transfers are not limited.

Just as there is no upper limit on the number of vehicle boardings, there is no lower limit either. Transit and walking are considered to effectively be a single mode. The practical implication of this is that the shortest path by "transit" is not required to include a transit vehicle. This may seem odd at first, but it allows the most consistent application and interpretation of the travel time calculation methodology. For example, the shortest walking path from an origin to a transit station often passes through destinations where job opportunities exist. In other cases, the shortest walking path from an origin to a destination might pass through a transit access point which provides no trips that would reduce the origin–destination travel time. In these situations, enforcing a minimum number of transit boardings would artificially inflate the shortest-path travel times. To avoid this unrealistic requirement, the transit travel times used in this analysis are allowed to include times achieved only by walking.

Transit accessibility is computed for every minute of the day, as described in Owen and Levinson (2015), which demonstrates that continuous accessibility metrics can provide a better description of the variation in transit commute mode share than do metrics evaluated at a single or optimal departure time.

Visualization

This project produces highly detailed accessibility datasets, and some level of aggregation is typically needed to produce easily understandable summary maps. Figures 4–7 provide examples of block-level accessibility results mapped at a constant data scale across four major metropolitan areas: Washington, DC; Atlanta, GA; Seattle, WA, and Minneapolis–Saint Paul, MN. In these maps, accessibility for each Census block has been averaged over the 7–9 AM period. The resulting average accessibility value indicates the number of jobs that a resident of each block could expect to be able to reach given a randomly-selected departure time between 7 and 9 AM.

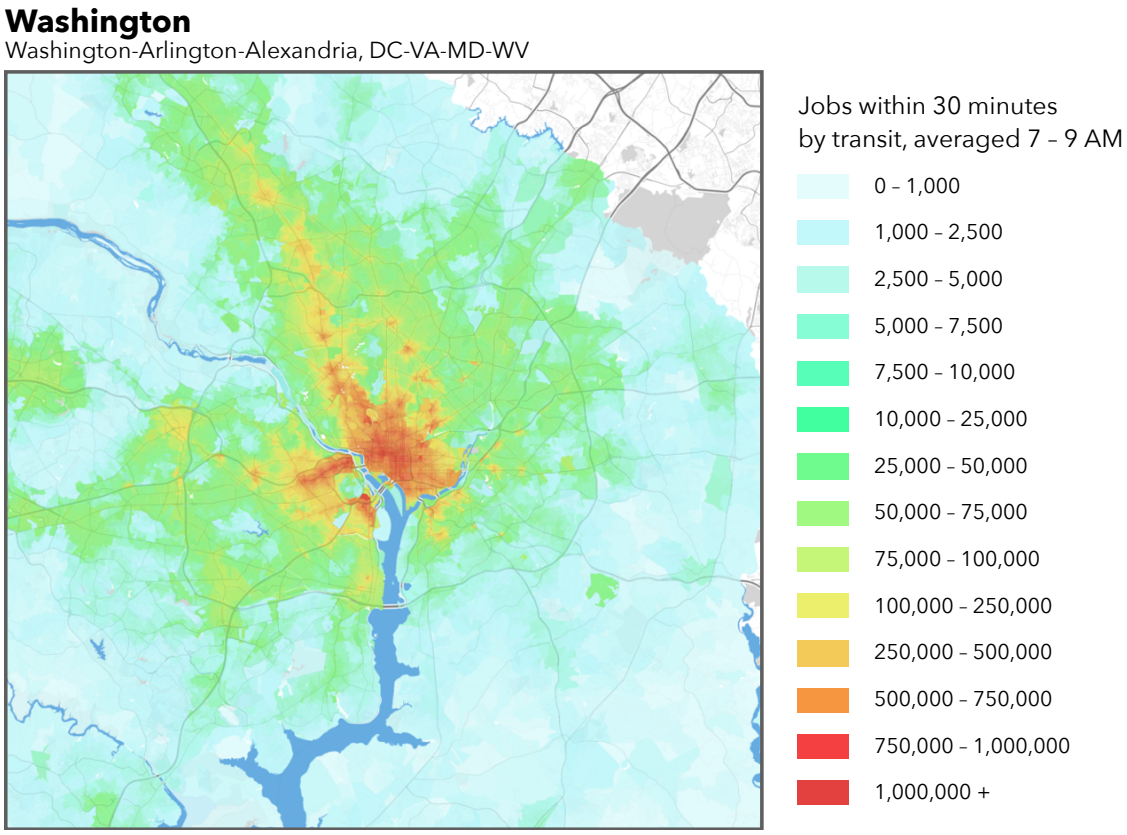

Figure 4: Map of job accessibility by transit in the Washington, DC metropolitan area.

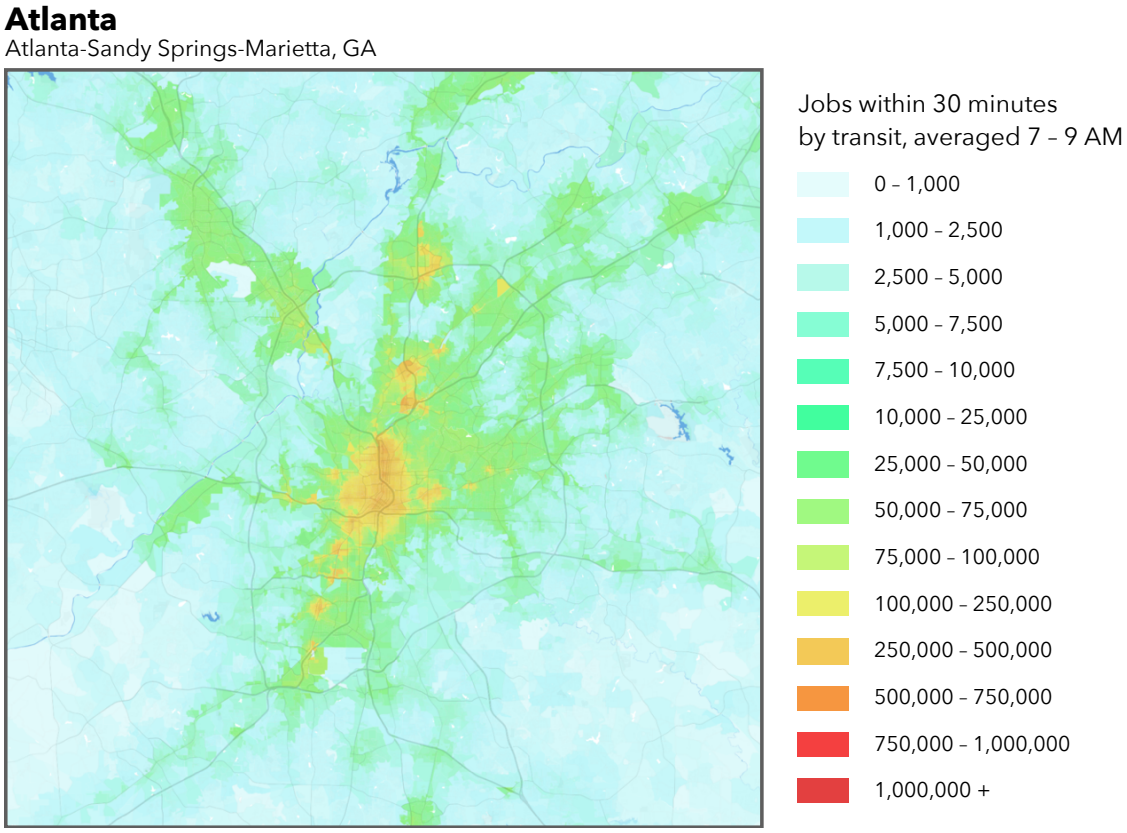

Figure 5: Map of job accessibility by transit in the Atlanta, GA metropolitan area.

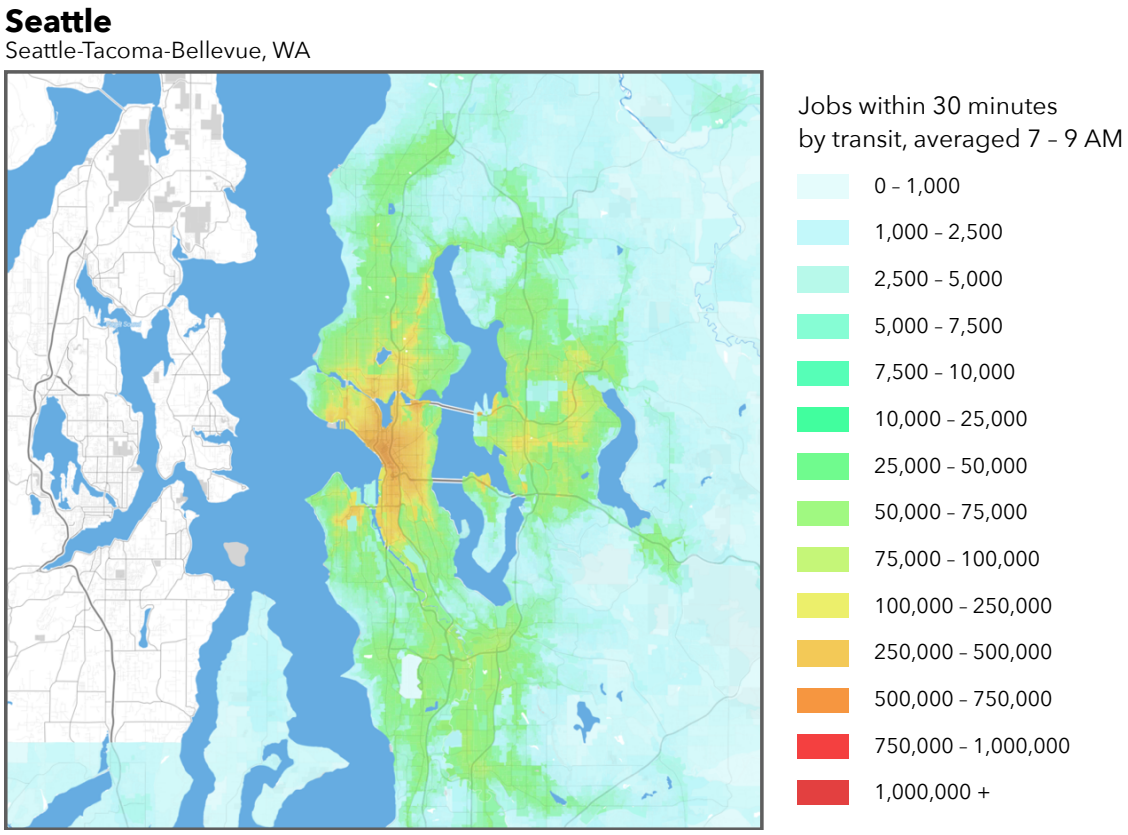

Figure 6: Map of job accessibility by transit in the Seattle, WA metropolitan area

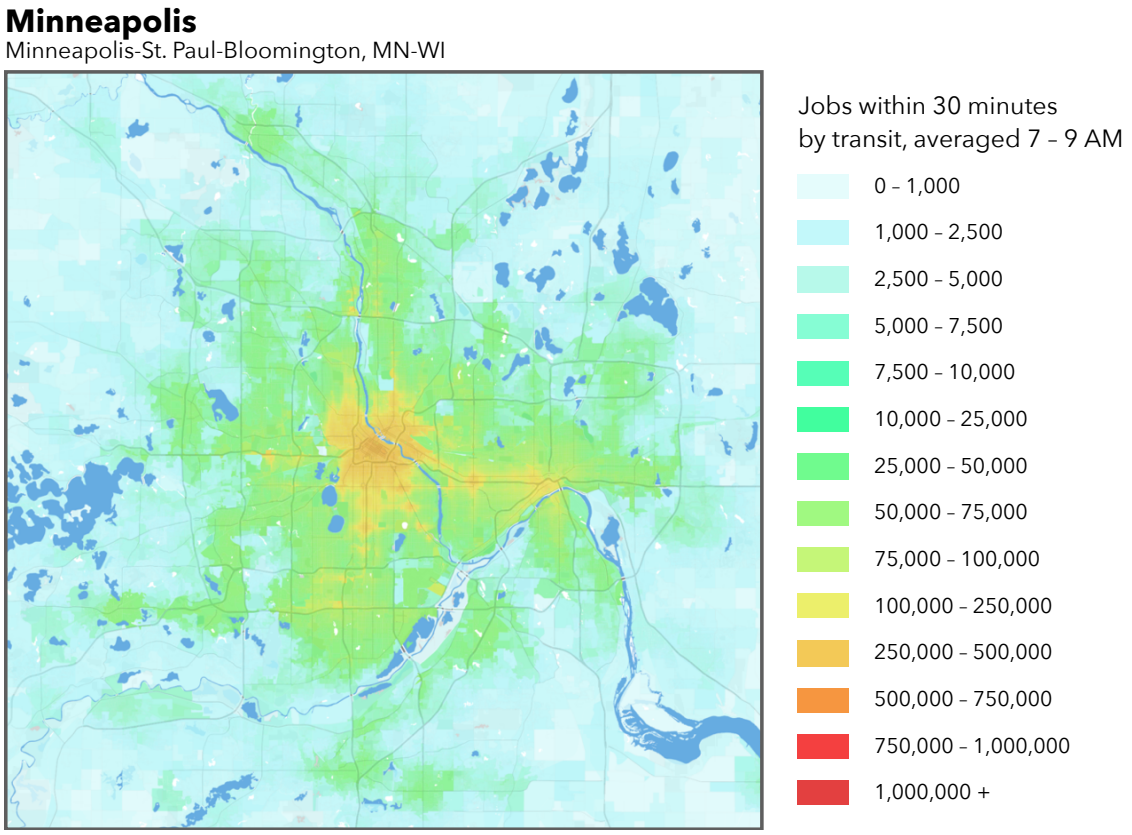

Figure 7: Map of job accessibility by transit in the Minneapolis–Saint Paul, MN metropolitan area

Conclusion

With the framework developed in this project, it is possible to evaluate the accessibility provided by public transit in any area where data is available. Within the United States, the only data limitation is the availability of transit schedules in GTFS format — all other sources are available with full national coverage (with the exception of LEHD data, which is not available for the state of Massachusetts). Also significantly, all data is public or available under an open license.

While this project adopted a specific accessibility metric (cumulative opportunities to jobs) and a set of parameters for implementing it, the framework itself provides flexibility. The core OpenTripPlanner software can calculate weighted accessibility; using a different destination dataset is a trivial modification; various travel time calculation parameters can be easily adjusted. While it is hoped that the accessibility data products described here will be useful for both research and practice, the framework can be used to fit a wide variety of specific accessibility evaluation scenarios. Consistency does not have to mean, "one size fits all."

This project also suggests ways that accessibility evaluation for other transportation modes could be improved. In some ways public transit is the most difficult domain

in which to perform this level of evaluation. Accessibility evaluations for car travel, for example, can employ the simplification of using average road speeds to avoid the need to calculate at multiple departure times with fewer consequences; network structure also remains constant over the course of the day. Given appropriate data sources, accessibility by car could be calculated for the same block-level resolution at a fraction of the computation costs.

However, this highlights a critical uniqueness of the transit case: travel time data (in the form of schedules) is publicly available. Outside of loop detector-based systems on urban highways (whose data format varies across cities and states), there exists virtually no equivalent for car travel. Open data initiatives in this realm, such as OpenTraffic.org, though promising, are nascent and lack coverage. Comprehensive data sources for road and highway speeds are effectively limited to commercial datasets; efforts to implement a similar evaluation for accessibility by car will need to confront this reality.

Acknowledgements

The project described in this article was sponsored by the University of Minnesota's Center for Transportation Studies. Many of the employed tools and methodological approaches were developed during earlier projects sponsored by the Minnesota Department of Transportation.

References

Catala, M., Downing, S. and Hayward, D. (2011). Expanding the Google transit feed specification to support operations and planning. Technical Report BDK85 997-15, Florida Department of Transportation.

Delling, D., Pajor, T., & Werneck, R. F. (2014). Round-based public transit routing. *Transportation Science*.

Google, Inc. (2013). *General transit feed specification reference*. [Online] Available from: https://developers.google.com/transit/gtfs/reference.

Jariyasunant, J., Mai, E., & Sengupta, R. (2011). Algorithm for finding optimal paths in a public transit network with real-time data. *Transportation Research Record: Journal of the Transportation Research Board*, (2256), 34-42.

Krizek, K., El-Geneidy, A., Iacono, M. and Horning, J. (2007). Refining methods for calculating non-auto travel times. Technical Report 2007-24, Minnesota Department of Transportation.

Krizek, K., Iacono, M., El-Geneidy, A., Liao, C.-F. and Johns, R. (2009). Application of accessibility measures for non-auto travel modes. Technical Report 2009-24, Minnesota Department of Transportation.

Levine, J., Grengs, J., Shen, Q., & Shen, Q. (2012). Does accessibility require density or speed? A comparison of fast versus close in getting where you want to go in US metropolitan regions. *Journal of the American Planning Association*, *78*(2), 157-172.

Owen, A. and Levinson, D. (2012). Annual accessibility measure for the Twin Cities metropolitan area. Technical Report 2012-34, Minnesota Department of Transportation.

Owen, A., & Levinson, D. M. (2015). Modeling the commute mode share of transit using continuous accessibility to jobs. *Transportation Research Part A: Policy and Practice*, *74*, 110-122.

Puchalsky, C. M., Joshi, D., & Scherr, W. (2012). Development of a regional forecasting model based on Google transit feed. In *91st Annual Meeting of the Transportation Research Board,* Washington, DC.

Wong, J. (2013). Leveraging the general transit feed specification for efficient transit analysis. *Transportation Research Record: Journal of the Transportation Research Board*, (2338), 11-19.